DOCUMENT RESUME

ED 285 918 TM 870 536

AUTHOR McBride, James R.
TITLE A Computerized Adaptive Edition of the Differential Aptitude Tests.
PUB DATE Aug 86
NOTE 15p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - General (140)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Identification; *Adaptive Testing; *Aptitude Tests; Comparative Testing; *Computer Assisted Testing; Factor Structure; Latent Trait Theory; Occupational Tests; Pretesting; Secondary Education; *Test Construction; Test Reliability; Test Validity; Timed Tests; *Vocational Aptitude
IDENTIFIERS *Differential Aptitude Tests; Rasch Model; Test Equivalence

ABSTRACT

An overview of the development of a computerized version of the Differential Aptitude Tests (DAT) is presented. It describes the previously existing printed version of the DAT, design of the computerized adaptive edition, calibration of the test items for use in the computerized version, and two field studies that compared the Adaptive and paper-and-pencil DATs to assess the reliability and validity of the adaptive edition. The DAT consists of seven power tests (Verbal Reasoning, Numerical Ability, Abstract Reasoning, Mechanical Reasoning, Space Relations, Spelling, and Language Usage) and one speeded test (Clerical Speed and Ability, CSA). The computerized edition, designed for use on the Apple II series, is half the length of the printed edition. Items were selected for the adaptive edition using a three-parameter item-response model, as well as the Rasch model. Results of the field tests showed that correlations between printed and computer editions were .78 to .88 on all seven power tests but .33 on the speeded test (CSA). No attempt was made to equate the adaptive CSA test with the printed one because of this low correlation. Factor loadings were found to be similar. It was concluded that the power tests could be considered psychometrically equivalent to their printed counterparts. (JGL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED285918

A Computerized Adaptive Edition of the Differential Aptitude Tests

James R. McBride
The Psychological Corporation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. McBride

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Convention of the American Psychological Association, Washington, D.C., August, 1986.

Copyright 1986, The Psychological Corporation

TM 870 536

A Computerized Adaptive Edition of the Differential Aptitude Tests

James R. McBride
The Psychological Corporation

Introduction

This presentation represents the unveiling of the first commercially published computerized adaptive test: the Computerized Adaptive Edition of the Differential Aptitude Tests, which is about to be published by TPC.

Since 1947, The Psychological Corporation has published the Differential Aptitude Tests, a battery of eight tests used for educational placement and for vocational guidance counseling, primarily in junior and senior high schools.

The Differential Aptitude Tests (DAT) have been revised a number of times over the years, and are highly regarded for their usefulness and technical quality. This occasion marks the introduction of a computer-administered, adaptive version of the DAT.

The Adaptive DAT is administered using Apple // series computers. The Apple // was chosen because it is the microcomputer most frequently used in the schools in 1986; almost 60 per cent of the computers in the schools are Apple //s.

The Adaptive DAT computer software is capable of administering all 8 DAT subtests, and the optional Career Planning Questionnaire. It scores each test immediately, and is capable of providing immediate results to the user, either as scores displayed on the computer screen, or as printed reports. The printed reports include an individual profile of norm-referenced (percentile) scores, and a narrative Career Planning Report (for those examinees who answered the Career Planning Questionnaire.)

All seven power tests of the DAT are administered adaptively; as a consequence of adaptive administration, they are only half the length of their printed counterparts. The eighth test -- Clerical Speed and Ability -- is a highly speeded test; the computer times its administration, probably more accurately that it is timed in typical classroom testing with the printed edition.

The reduced length of the 7 adaptive tests makes the Adaptive DAT considerably more efficient than the printed edition. The printed edition typically takes about three and a half hours to administer; when it is given in short class periods, it requires six sessions to give the complete battery. In contrast, the Adaptive DAT takes an average of about 90 minutes, or 3 class periods, to administer. It is self-paced, however, so that the average time reflects some cases in which it takes less than an hour, and others in which it takes more than 2 hours.

This presentation is an overview of the development of the Adaptive DAT. It will describe the design of the adaptive edition, the calibration of DAT test items for

use in the adaptive edition, and the empirical research that has been conducted to compare the Adaptive DAT with its printed counterpart.

The presentation is organized into five sections. The first section briefly describes the current, printed edition of the Differential Aptitude Tests. The second section describes the design of the computerized adaptive edition, and the source and calibration of test items used in it. The third section describes the empirical research into its reliability and validity. The fourth section presents the results. Finally, the fifth section provides a discussion and summary.

Section 1: The Differential Aptitude Tests

Some background on the DAT was presented in the introduction. This section will briefly give some additional information about the current edition of the printed DAT.

The DAT consists of eight tests, seven of which are essentially power tests, and one of which is a short speeded test. The test's names and standard abbreviations are listed below:

VR	Verbal Reasoning	50 5-choice items
NA	Numerical Ability	40 5-choice items
AR	Abstract Reasoning	45 5-choice items
CSA	Clerical Speed and Ability	100 5-choice items
MR	Mechanical Reasoning	70 3-choice items
SR	Space Relations	60 4-choice items
SP	Spelling	90 2-choice items
LU	Language Usage	50 5-choice items

All seven of the power tests are timed; however, the time limits are fairly generous. Of the seven, the first five are primarily aptitude tests; the last two, Spelling and Language Usage, are best described as achievement tests. The speeded test is Clerical Speed and Ability; it is administered in two parts, the first of which is unscored and constitutes practice for the second part.

Scores on the DAT are reported both as raw scores and as percentile ranks. There are separate percentile norms for grades 8, 9, 10, 11, and 12. In addition, there are separate norms for males and females, and for Fall and Spring school semesters.

Section 2: Design of the Computerized Adaptive Edition

The "design" of a computerized adaptive test encompasses several important technical issues: First is the broad issue of the general technical approach to

take. Most recent adaptive test development has employed item response theory; the Adaptive DAT is no exception.

Given that the test will be based on item response theory, the second issue is the choice of item response model. The leading candidates are the 3-parameter logistic model, the similar 3-parameter normal ogive model, and the 1-parameter logistic model known as the Rasch model.

The third issue concerns the methods to be used for item selection and for ability estimation. The fourth issue is the matter of what criterion to use to terminate the test.

The balance of this section will be a discussion of the decisions made regarding each of these issues in the design of the Adaptive DAT.

The General Approach

From the outset, item response theory was the intended technical foundation of the design of the Adaptive DAT. The choice of IRT was not difficult, given the author's own experience, and the burgeoning number of applications of IRT both to adaptive testing and to conventional test design.

Choice of Item Response Model

Of the three models cited above, two have been used most often in adaptive testing research and development. The 3-parameter logistic (3-PL) model has been used most often, in such applications as the Army's CAST (Computerized Adaptive Screening Test), the Navy's CAT/ASVAB system, and the College Board's Computerized Placement Tests. The 3-parameter normal ogive model has been used less extensively, but its use in adaptive testing has been exhaustively explicated by Urry. The Rasch model, although seemingly inappropriate for use with multiple choice test items, has also been used successfully in adaptive testing (see, for example, Moreno, Wetzel, McBride & Weiss (1984), McBride & Moe (1986), and McKinley & Reckase (1980)).

Practical considerations led to the adoption of the Rasch model as a temporary measure early in the development of the Adaptive DAT. However, from the beginning there was a plan to replace the Rasch model with the more general 3-parameter logistic model at the earliest opportunity, unless evidence indicated that the change of model was unwarranted. The normal ogive model was not seriously considered because of the absence of well-developed software for fitting it to data.

An early step, after settling on the Rasch model as an interim measure, was to develop the Adaptive DAT's item banks. Since there was a large and readily available body of item response data from the printed DAT, one of the two forms of the printed DAT, Form V, was chosen to provide the items for the adaptive test battery. An item and test analysis program called TRIAN was used to fit the

Rasch model to item response data sampled from the 1982 standardization of DAT forms V and W. (Later in the project, the 3-parameter model was applied to the standardization data, using the MicroCAT program (Assessment Systems Corporation, 1985)).

Item Selection and Ability Estimation Methods

The author's experience in the CAT/ASVAB project led to the choice of a hybrid approach to these two central adaptive testing functions. This hybrid approach uses Owen's Bayesian sequential updating technique to estimate the examinee's ability, and uses a modified "maximum information" criterion for item selection. (The modification consists of preparing item selection look-up tables that reflect item information at various ability levels. At each point in the test, either the "best" item is chosen, or a random choice is made from among the 2 to 5 best items.)

This hybrid approach has been shown in simulation studies (e.g. Wetzel & McBride, 1983) to be practically equivalent to the best alternative strategies. It has been used successfully, and very effectively, in the Department of Defense research and development of an adaptive version of the Armed Services Vocational Aptitude Battery, ASVAB.

Since the Differential Aptitude Tests are used to assess examinees ranging from the eighth through the twelfth grades, with grade-specific norms, it made sense to consider using different starting levels of difficulty, based on the examinee's grade level. This is done in the Adaptive DAT; the starting point for each subtest is predicated on the mean performance of that grade level in the 1982 standardization.

Test Termination Criterion

Each of the seven adaptive tests in the Adaptive DAT is a fixed length test. Each test terminates when the number of items administered is half the length of the counterpart printed DAT test.

Summary

To summarize the design features of the Adaptive DAT: the battery includes seven adaptive power tests, intended to be alternate "forms" of the printed edition tests. Each test is individually administered by choosing IRT-calibrated items from a bank consisting of all the items in Form V of the printed edition. All of the items have been calibrated using the Rasch model. Each adaptive test uses Owen's Bayesian sequential updating procedure to estimate the examinee's ability after answering each question. In each case, the Bayes prior mean is the mean performance of the examinee's grade level in the standardization of the printed DAT. Once the ability estimate is updated, a modified maximum information item selection procedure is used. Each adaptive test terminates

when its length is half the number of items in the counterpart test in the printed DAT.

Section 3: Empirical Research

The Computerized Adaptive Edition is intended to be used interchangably with the printed forms of the Differential Aptitude Tests. For ease of interpretation of test results, this made it desirable that the adaptive test use the same norms as Forms V and W of the printed edition. To justify this, it was necessary to establish a high degree of correspondence between the adaptive and the printed editions, and then to equate the adaptive DAT test scores with the printed test.

Establishing the correspondence of the two different modes of administration meant demonstrating that the two correlated highly, and had similar factorial structures. Equating the two meant deriving transformations that would permit expressing the adaptive test scores as equivalent raw scores of the printed edition.

To accomplish these two purposes, two field tests were conducted, one in the Fall of 1985, and one in the Spring of 1986. These two field tests and their purposes are described in the remainder of this section.

Field Test #1: Fall 1985

In the initial field test of the Adaptive DAT, examinees were to take the entire DAT test battery in two modes: computer-administered, and conventionally administered. Since all the items in the computerized adaptive edition are taken from Form V, Form W was used for the conventionally administered tests. Order of administration was counterbalanced.

Twelve school districts around the nation participated in this field test; in all, over 500 examinees -- most of them students in grades 8 through 12 -- took part. Of these, approximately 425 completed both the adaptive and the printed editions.

The correlation matrix of the DAT scores -- including both the adaptive and the printed edition scores -- were subjected to factor analysis to assess the degree to which the two different modes of administration measured the same latent variables. For scale comparability, the printed Form W raw scores were transformed to Rasch ability estimates, using the same metric as the adaptive test scores.

The factor analysis extracted four factors, using the principal axes method and Varimax rotation to simple structure.

Field Test #2: Spring 1986

The design of the Spring field test was both similar to and different from the Fall one. The similarity lay in the fact that examinees were administered both computerized adaptive and printed DAT tests, in counterbalanced order. The principal difference was that in the Spring field test, the battery was divided into three parts, and each examinee took only one part.

The three parts consisted of the following subtests:

Part 1	Verbal Reasoning
	Numerical Ability
Part 2	Abstract Reasoning
	Clerical Speed and Ability
	Spelling
	Language Usage
Part 3	Mechanical Reasoning
	Space Relations

Dividing the battery into these three parts was intended to increase participation in the research by reducing the amount of examinee time required. In the Fall field test, each examinee was tested for a total of about nine class periods; each of the parts in the Spring field test was designed so that the computerized adaptive version could be administered in one class period, and the printed version could be administered in two class sessions. Thus, each examinee could complete the Spring tests in three class periods.

To aid in counterbalancing the order of administration, each part consisted of two alternate forms, designed in such a way as to assure that administration of the computerized version of any DAT subtest would be preceded by the printed version in about half the cases.

Almost 4,000 examinees, from a total of 90 geographically distributed school districts, participated in the Spring field test. Some, however, did not take both versions of the tests. About 2700 examinees had complete data for both versions; these were distributed fairly evenly among Parts 1, 2, and 3.

Correlations Across Modes of Administration

To assess the degree of correspondence of the computerized adaptive DAT tests with their printed counterparts, correlations of test scores attained in the two modes of administration were calculated. These correlations may be interpreted as alternate form correlations, although the forms are not strictly parallel.

Table 2 contains these alternate test correlations, for the group as a whole and for each of the five grade levels present in the study. For all subtests except

Clerical Speed and Ability, the correlations are between actual Form W scores, and adaptive test scores transformed into Form W equivalent scores. The data in Table 2 will be discussed in Section 4, below.

Equating the Computerized Adaptive and Printed Tests

In order to use the 1982 DAT norms for reporting and interpreting examinee performance on the Adaptive DAT, it was essential to equate the adaptive test to the printed one. Four different approaches to this equating were tried: linear equating, equipercentile equating, and two different IRT equating procedures. Before equating transformations were derived, the adaptive test protocols were rescored using the 3-parameter logistic item parameters, so there were two sets of adaptive test scores: one consisting of Rasch ability estimates, the other consisting of 3-parameter model ability estimates.

Since there were seven adaptive tests being equated, there were a total of 56 separate equatings (seven tests times four methods, times two ability estimates per method). Both linear and equipercentile equating were consistently advantageous compared to the IRT approaches, so the latter will not be discussed further. For both the linear and the equipercentile equating, the Spring field test sample was partitioned into a derivation subsample and a cross-validation or evaluation subsample. The derivation subsamples numbered 700 examinees for all tests except Mechanical Reasoning and Space Relations. For the latter two tests, a slightly smaller derivation sample was used.

The evaluation subsamples consisted of the 200 to 250 Spring field test cases not used to derive the equating transformations. The Fall field test samples were also used, as separate samples, to evaluate the quality of the equating. Several criteria were evaluated, including the following, in the evaluation samples:

- Correlations of transformed scores with Form W obtained raw scores;
- Mean discrepancies between transformed and obtained raw scores;
- Standard deviations of discrepancies;
- Mean squared discrepancies.

Section 4: Results

Correlation Analyses

The intercorrelations of the adaptive test scores with the printed Form W scores are displayed in Table 1.

Table 1 here.

The table includes correlations for Grades 8 through 12 separately and combined. For the combined grades, of the seven adaptive tests (i.e., every test except Clerical Speed and Ability) the correlations across mode ranged from .78 to .88, with a median correlation of .85 . The highest correlations were for the Verbal Reasoning, Numerical Ability, Spelling, and Language Usage tests, for which all correlations were .85 or higher. The lowest adaptive test correlations were those of the three pictorial tests: Mechanical Reasoning, Space Relations, and Abstract Reasoning; their correlations ranged from .78 to .82.

By far the lowest correlation across modes of administration was observed for the Clerical Speed and Ability test, where the correlation was .33 .

Factor Structures

The results of the factor analysis of the Fall field test data are displayed in the rotated factor matrix in Table 2.

Table 2 here.

Factor 1 in this matrix can be interpreted as a "verbal information" factor. The Spelling tests of both modes have the highest loadings, followed by both Language Usage tests; the Spelling and Language Usage tests both require the examinee to recognize errors in previously learned verbal material. The Verbal Reasoning and Numerical Ability tests have smaller, but nonetheless strong, loadings on this factor. The loadings of these salient subtests on Factor 1 were nearly identical across mode of administration.

Factor 2 seems to be a "non-verbal figural reasoning" factor. The strongest loadings were for the Abstract Reasoning and Space Relations subtests, both of which involve reasoning tasks using unfamiliar figural stimuli. As with Factor 1, the loadings of a named subtest were very similar, regardless of mode of administration.

Factor 3 is effectively defined by the Mechanical Reasoning subtests in both modes of administration. Similarly, the two Clerical Speed and Ability tests mark Factor 4.

For all four factors, the patterns of factor loadings were very similar for both the computerized and the printed versions of the tests.

Equating

The equating data will not be presented here; they will be the subject of a forthcoming separate paper. Two items should be noted in passing, however. The first is that no attempt was made to "equate" the computer-administered Clerical Speed and Ability test with the printed one, because of the low correlation between them; instead, norms based on computer-administration were derived.

The second noteworthy item is that there was no appreciable advantage of the ability estimates based on the 3-parameter logistic model, over the Rasch model ability estimates. Scores derived from both models had nearly identical levels of correlations with obtained Form W scores, and yielded similar levels of equating error.

Section 5: Summary and Discussion

The results of the two field tests show a high degree of correspondence between the computerized adaptive DAT tests and their printed Form W counterparts, but little correspondence between the computerized and the printed Clerical Speed and Ability (CSA) tests. This discussion will deal first with the seven adaptive power tests, and last with the CSA tests.

The correlations of the seven power tests across modes of administration were high enough to consider the computerized adaptive and the printed editions of the DAT as alternate -- but of course not parallel -- tests. This is supported by the results of the factor analysis, which show a very high degree of similarity of the two batteries, in terms of their patterns of factor loadings. For practical purposes, the patterns of factor loadings of the computerized tests and the printed tests were identical. (Even the computerized CSA test had its highest loading on the clerical speed factor.)

Given its high degree of correspondence with the printed edition tests, the Computerized Adaptive Edition could be considered psychometrically equivalent to it, except for the very different metric used in the adaptive tests. The equating was intended to remedy this, by transforming the adaptive test scores into the raw score metric of the printed test. Although the results were not reported here, the equating results were quite satisfactory; they will be reported on a subsequent occasion.

The two different modes of administering the CSA test, however, did not correlate highly enough to justify equating. Both the correlation analysis and the factor analysis indicate that the computerized CSA test is measuring a somewhat different variable than the printed version. Additional results, not reported here, bear this out. The low correlation between the two CSA tests cannot be attributed to content differences, because the items are identical except for order. The difference probably lies in the different tasks involved in responding

to CSA items on the computer screen rather than on an answer sheet. More research is needed into the explanation of these observed differences, and their implications for predicting examinee behavior.

The development and research into the Computerized Adaptive Edition of the DAT will be fully reported in a forthcoming technical report of The Psychological Corporation. That report will be in the form of a supplement to the technical material on the printed editions of the DAT. It will be intended to address the documentation requirements of both the Standards for Educational and Psychological Testing and the APA Guidelines for Computer-Based Tests and Interpretations.

REFERENCES

Assessment Systems Corporation (1985). User's Manual for the MicroCAT Testing System. (Research Report 85-1). St. Paul, Minnesota, June 1985.

McBride, J.R. & Moe, K.C. (1986). . Paper presented at the Annual Convention of the National Council on Measurement in Education. San Francisco, April 1986.

McKinley, R. & Reckase, M.D. (1980). A successful aplication of latent trait theory to tailored testing. (Research Report 80-1). Columbia, Missouri: Educational Psychology Department, Universoty of Missouri.

Moreno, K.E., Wetzel, C.D., McBride, J.R. & Weiss, D.J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Wetzel, C.D. & McBride, J.R. (1983). Influence of fallible item parameters on test information during adaptive testing. (NPRDC TR 83-15). San Diego, CA: Navy Personnel Research and Development Center, April 1983.

Table 1: Correlations Between Computerized Adaptive and Printed (Form W) Tests of the Differential Aptitude Tests

	Grade						N
Test Name	8	9	10	11	12	All	
Verbal Reasoning	84	89	88	85	89	88	942
Numerical Ability	86	87	87	80	84	86	924
Abstract Reasoning	82	79	82	84	84	82	949
Clerical Speed and Ability	26	28	38	32	25	33	948
Mechanical Reasoning	78	79	79	71	79	78	829
Space Relations	80	87	86	76	75	81	828
Spelling	82	89	85	86	87	86	960
Language Usage	85	85	85	83	88	85	959

(Decimal points omitted from correlations.)

Table 2: Varimax Rotated Factor Matrix
of the Computerized Adaptive and Printed Editions
of the Differential Aptitude Tests (N = 350).

	Factor			
	I	II	III	IV
Computer-administered Tests				
Verbal Reasoning	72	37	31	12
Numerical Ability	60	59	09	17
Abstract Reasoning	36	74	16	12
Clerical Speed and Ability	32	32	- 02	68
Mechanical Reasoning	19	24	86	11
Space Relations	21	82	23	06
Spelling	88	13	04	14
Language Usage	79	29	21	07
Printed Tests, Form W				
Verbal Reasoning	70	41	36	14
Numerical Ability	60	58	05	26
Abstract Reasoning	33	73	26	08
Clerical Speed and Ability	14	-01	24	86
Mechanical Reasoning	16	40	80	15
Space Relations	12	77	43	06
Spelling	88	12	02	21
Language Usage	81	29	18	14

(Decimal points omitted.)